Math 118: Mathematical Methods of Data Theory

Lecture 9: Graphs and Spectral Clustering

Instructor: Daniel Mckenzie

Dept. of Mathematics, UCLA

TBD

Graphs

- Graphs $G = (V, E)$ where $V =$ vertex set and $E =$ edge set.
- For this class $V = \{v_1, \dots, v_n\}$ and write (i, j) for edge between v_i and v_j.
- **Adjacency matrix:** $A \in \mathbb{R}^{n \times n}$ with $A_{ij} = 1$ if (i, j) is edge, and $A_{ij} = 0$ otherwise.

Insert Adjacency matrix and small graph here

Graphs

- $d_i =$ degree of v_i $=$ number of edges incident to v_i.
- $D = \text{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n \times n}$.
- The graph Laplacian: $L = D - A$.
- Important properties of L:
  - L is symmetric and pos. semi-definite.
  - $L\mathbf{1} = \mathbf{0}$.
- *Further variants:* G can have *weighted* or *directed* edges.

Examples of Graphs

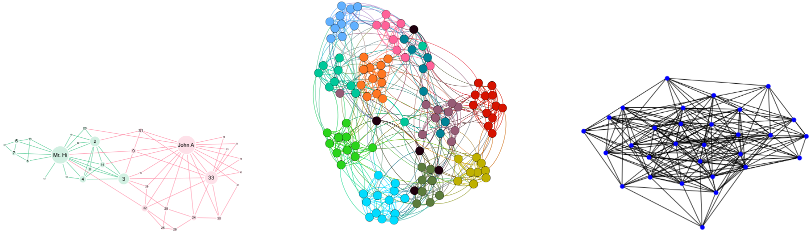

Figure: Left to right: Zachary's Karate club[3], College Football 2000 season [4], Erdos-Renyi random graph generated using `networkx`

Graphs often called *networks* in applied settings.

[1] Originally: *An information flow model for conflict and fission in small groups* Zachary, W. 1977. Image from https://studentwork.prattsi.org/infovis/labs/zacharys-karate-club/

[2] Originally: *Community structure in social and biological networks.* Girvan & Newman (2002). Image from *Compressive sensing for cut improvement and local clustering* Lai & Mckenzie (2020)

[3] Originally: *An information flow model for conflict and fission in small groups* Zachary, W. 1977. Image from https://studentwork.prattsi.org/infovis/labs/zacharys-karate-club/

[4] Originally: *Community structure in social and biological networks.* Girvan & Newman (2002). Image from *Compressive sensing for cut improvement and local clustering* Lai & Mckenzie (2020)

Connected Components and Clusters

- C_1 is a **connected component** of G if no edges between C_1 and $V \setminus C_1$.
- **Corollary:** If C_1 is a connected component then so is $C_2 = V \setminus C_1$.

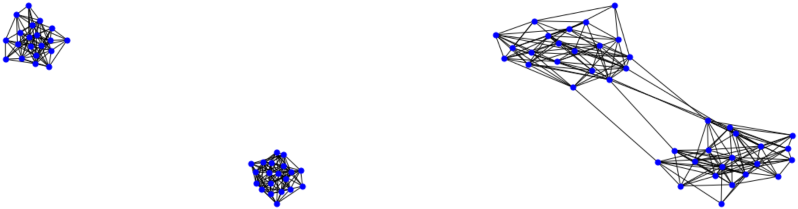

Figure: Left: Two connected components. Right: One connected component but two clusters

- C_1 is a **cluster** of G if "few" edges between C_1 and $V \setminus C_1$ **and** many internal edges in C_1.
- Ratio Cut.
  - Let $e(S, V \setminus S) = \#$ edges from S to $V \setminus S$.
  - $\text{RCut}(S) = \dfrac{e(S, V \setminus S)}{|S||V \setminus S|}$.
  - Find cluster as $C = \arg\min_{S \subset V} \text{RCut}(S)$.

Why is finding clusters hard?

- Finding connected components: Breadth-First Search or Depth-First Search.
- Min Cut:
  - Recall $e(S, V \setminus S) = \#$ edges from S to $V \setminus S$.
  - Min Cut problem: Find $C = \underset{S \subset V}{\arg\min}\, e(S, V \setminus S)$.
  - Can be done efficiently $(O(n^3))$ using Ford-Fulkerson algorithm.
  - **Problem:** typically finds small C.
- Recall $\mathrm{RCut}(S) = \dfrac{e(S, V \setminus S)}{|S||V \setminus S|}$.
- Unfortunately $C = \underset{S \subset V}{\arg\min}\, \mathrm{RCut}(S)$ is NP-hard.
- Thus, resort to approximate algorithms, like Spectral Clustering.

The Spectral Clustering Algorithm

- Spectral clustering for 2 clusters:
  1. Compute d_i for $i = 1, \ldots, n$. Let $D = \mathrm{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n \times n}$.
  2. Compute Laplacian: $L = D - A$.
  3. Compute **second** eigenpair $(\lambda_2, \boldsymbol{v}_2)$.
  4. Assign vertices to clusters as:

  $$v_i \in C \text{ if } (\mathbf{v}_2)_i > 0 \text{ or } v_i \in V \setminus C \text{ if } (\mathbf{v}_2)_i < 0$$

  5. **Output:** C.

Output of Spectral Clustering

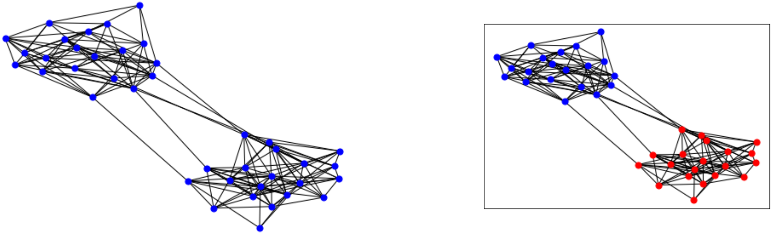

Figure: Left: Two connected components. Right: One connected component but two clusters

Analysis of Spectral Clustering

- Recall solving $C = \arg\min_{S \subset V} \left\{ \mathrm{RCut}(S) = \dfrac{e(S, V \setminus S)}{|S||V \setminus S|} \right\}$. is NP-hard.
- Instead, will show that Spectral Clustering solves a *relaxed version* of Ratio Cut.
- Proceed via steps:
  1. Introduce indicator vectors $\mathbf{1}_S \in \mathbb{R}^n$ for $S \subset V$.
  2. Relate to Ratio Cut: $\mathrm{Rcut}(S) = \frac{1}{n^2} \mathbf{I}_S^\top L \mathbf{I}_S$.
  3. **Relax:** Replace $\mathbf{1}_S \in \mathbb{R}^n$ with arbitrary $\boldsymbol{v} \in \mathbb{R}^n$.
  4. Argue that solving relaxed problem is easy: $\boldsymbol{v}_2 = \arg\min_{\boldsymbol{v} \in \mathbb{R}^n} \boldsymbol{v}^\top L \boldsymbol{v}$.
  5. Can (approximately) reconstruct C from $\boldsymbol{v}_2$.

Ensure consistency with notation and type of indicator vectors, add a small (4–6 vertex) running example. Check consistency between S and C.

Analysis of Spectral Clustering

- For any $S \subset V$ define: $\mathbf{I}_s = \begin{cases} \sqrt{\frac{|S^c|}{|S|}} & \text{if } v_i \in S \\ -\sqrt{\frac{|S|}{|S^c|}} & \text{if } v_i \notin S \end{cases}$
- Properties of indicator vectors:
  1. $\text{Rcut}(S) = \frac{1}{n^2}\mathbf{I}_S^\top L\mathbf{I}_S$ *(Homework)*.
  2. So: $C = \text{argmin}_{S \subset V}\text{RCut}(S) \Leftrightarrow I_C = \arg\min_{S \subset V} \mathbf{I}_S^\top L\mathbf{I}_S$.
  3. $\mathbf{1}^\top \mathbf{I}_S = 0$. Proof:

$$\mathbf{1}^\top \mathbf{I}_S = \sum_{i \in V} (\mathbf{I}_S)_i = \sum_{v_i \in S} \left(\sqrt{\frac{|S^c|}{|S|}}\right) + \sum_{v_i \in S^c} \left(-\sqrt{\frac{|S|}{|S^c|}}\right)$$
$$= |S| \left(\sqrt{\frac{|S^c|}{|S|}}\right) - |S^c| \left(\sqrt{\frac{|S|}{|S^c|}}\right)$$
$$= \sqrt{|S||S^c|} - \sqrt{|S||S^c|} = 0$$

  4. If $S \neq \emptyset, V$ then $\|\mathbf{I}_S\|_2 = \sqrt{n}$ *(Homework)*.

- Relax problem $\arg\min_{S \subset V} \mathbf{I}_S^\top L\mathbf{I}_S$ to $\arg\min_{\substack{\mathbf{v} \in \mathbb{R}^n \\ \|\mathbf{v}\|_2 = \sqrt{n} \text{ and } \mathbf{1}^\top \mathbf{v} = 0}} \mathbf{v}^\top L\mathbf{v}$

Analysis of Spectral Clustering

Need a detour on eigenvalues and Rayleigh-Ritz. Caution that now enumerating eigenvalues in *increasing* order.

- **Claim:** $\boldsymbol{v}_2 = \arg\min_{\boldsymbol{v}\in\mathbb{R}^n} \mathbf{v}^\top L\mathbf{v} : \ \mathbf{1}^\top\mathbf{v} = 0$ and $\|\mathbf{v}\|_2 = \sqrt{n}$. Why?
- First eigenvector: $\mathbf{1} = \boldsymbol{v}_1 = \underset{\substack{\boldsymbol{v}\in\mathbb{R}^n \\ \|\mathbf{v}\|_2=\sqrt{n}}}{\arg\min}\ \boldsymbol{v}^\top L\boldsymbol{v}$.
- Second eigenvector: $\boldsymbol{v}_2 = \underset{\substack{\boldsymbol{v}\in\mathbb{R}^n \\ \|\mathbf{v}\|_2=\sqrt{n} \text{ and } \mathbf{1}^\top\mathbf{v}=0}}{\arg\min} \mathbf{v}^\top L\mathbf{v}$
- So:

$$\mathbf{I}_C = \underset{S\subset V}{\arg\min}\, I_S^\top L I_S \approx \underset{\substack{\boldsymbol{v}\in\mathbb{R}^n \\ \|\mathbf{v}\|_2=\sqrt{n} \text{ and } \mathbf{1}^\top\mathbf{v}=0}}{\arg\min} \mathbf{v}^\top L\mathbf{v} = \boldsymbol{v}_2$$

- $(\mathbf{I}_C)_i > 0$ if $v_i \in C$ and $(\mathbf{I}_C)_i < 0$ if $v_i \notin C$.
- Use same rule with $\boldsymbol{v}_2$:

$$v_i \in C \text{ if } (\mathbf{v}_2)_i > 0 \text{ or } v_i \in V \setminus C \text{ if } (\mathbf{v}_2)_i < 0$$